Corey McCrea

CPSC 298 – Intro to R

5/14/19

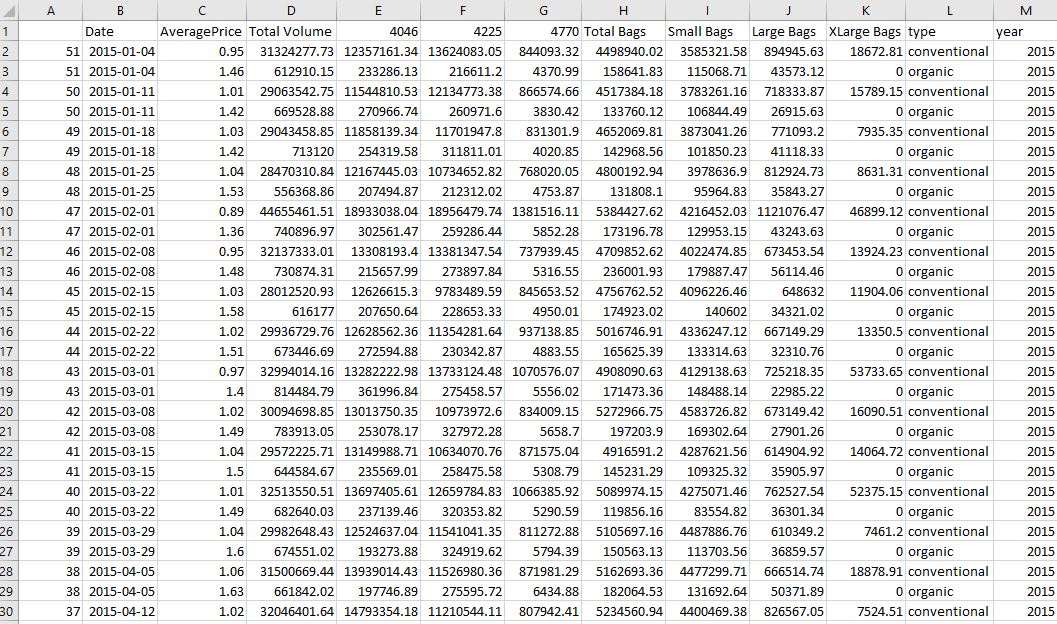
Avocado Analysis

Link to Slides: <https://bit.ly/presentMe>

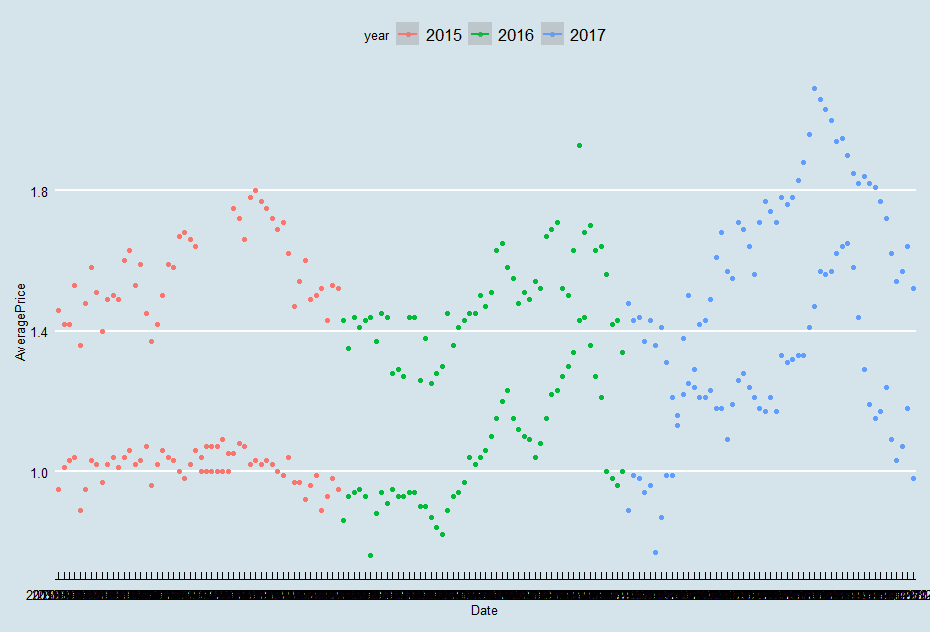
**Dataset/Summary**

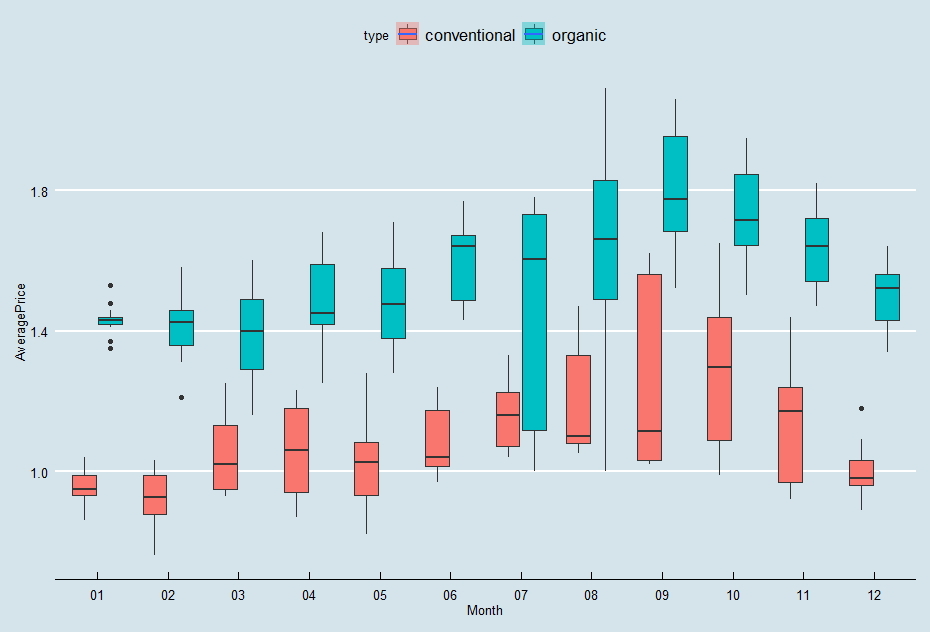
I found a [Kaggle dataset](https://www.kaggle.com/neuromusic/avocado-prices) on avocado pricing and volume over the years 2015-2018. It provides 12 stats (date, average price, total volume, type, year, region) regarding avocados over a three-year period (2015-2018). Date is the weekly date of observation. Average price is the average price of a single avocado that week. Type is whether the avocado is conventional or organic. Year is the year of observation. Region is the city or region that the avocados are being sold. Total volume is the number of avocados sold. 4046, 4225, 4770 are the different avocado types, i.e. small Haas etc.

In order to use this dataset effectively, I had to do some cleaning first. I only looked at the total US region and I ignored 2018 as there wasn’t a complete year of entries.

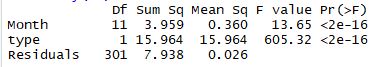


**Analysis/Graphs**

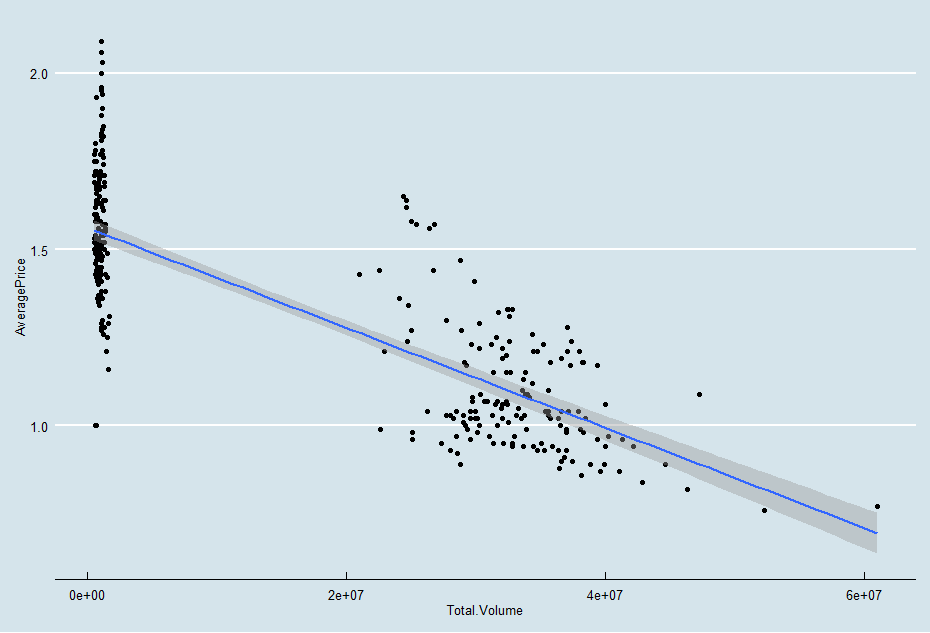
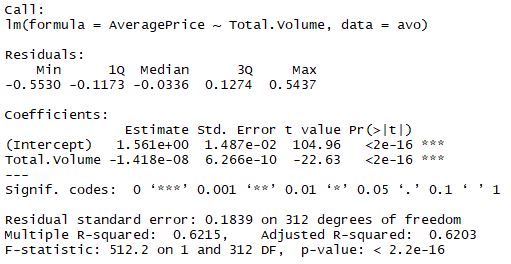
The first thing I looked at was average price and aate to get a general, overarching idea of the trends in this dataset.

Here a cyclical pattern can be seen going on throughout the years, which could be due to avocadoes going in season during the summer and then out of season. There is a spike in 2017 which could be explained as a result of the California fires impacting avocado groves as California is one of the largest avocado suppliers. We also see two groups the upper group, organic and the lower group conventional. This can be seen more clearly in this average price and month box plot; this contains data from all three years.

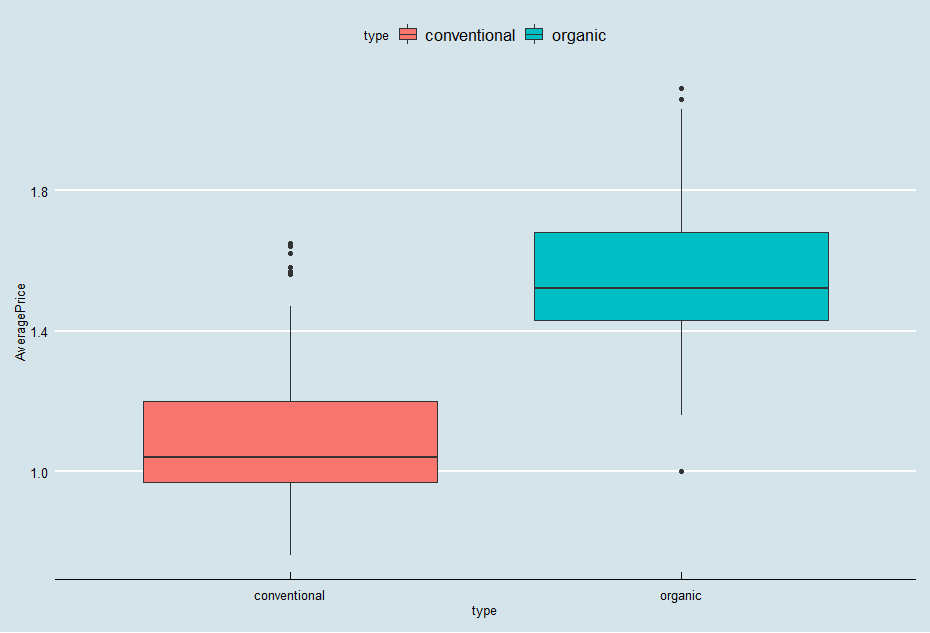
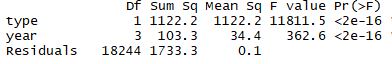
I sought to confirm this relationship with an **ANOVA** analyzing average price with month combined with type.

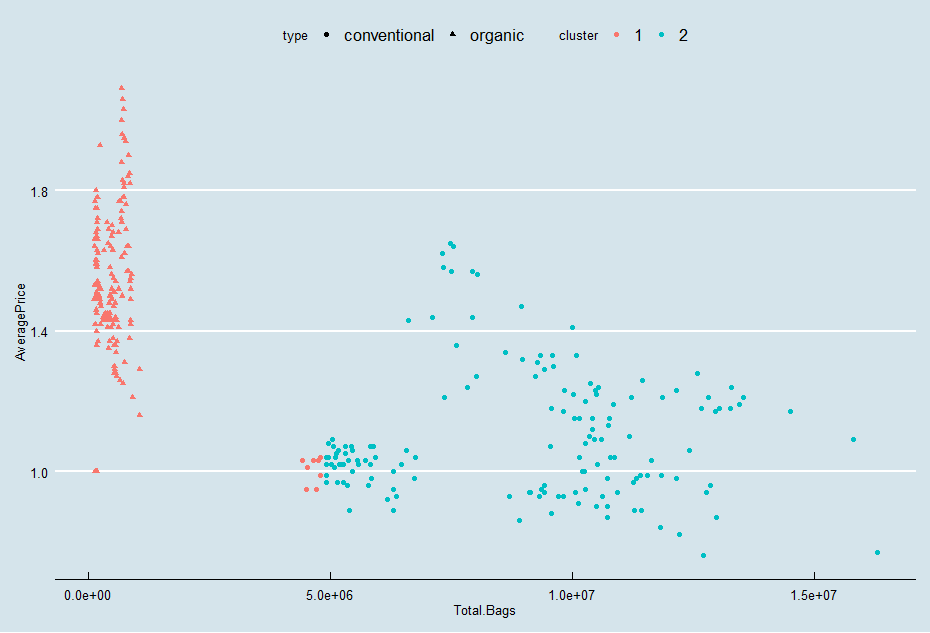
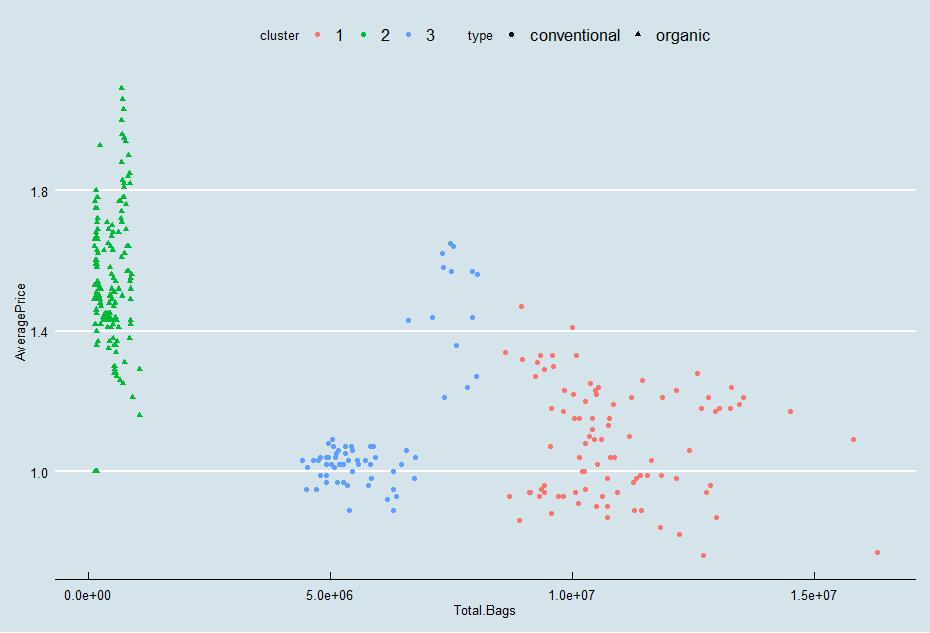


There was a significant relationship between these variables as our p-value is < .05. This is evident from the above boxplot.

I then ran a **regression model** to explore the expected relationship between price and volume.

Due to a p-value of <.05 there is a significant relationship between price and volume as one would expect. It is clearly a negative relationship as is seen in the graph, as volume goes up the price goes down. From a business standpoint this is very logical, supply and demand are directly related.

I then wanted to confirm that price and type of avocado were related so I ran another **ANOVA**. I found that there is a significant relationship between these two. Organic avocado prices average a whole 50 cents higher than conventional avocadoes.

For my fourth analysis I ran a **k-means clustering** starting with k =2 to cluster sales into two groups based on total bags and average price. The clustered groups were very close to being separated by type, so I then increase the k value to three and one group was organic and two were conventional. If I had more time I would have enjoyed exploring if there was a relationship between the two conventional groups.

https://lh4.googleusercontent.com/mjbVcoPD-evE9EZTzhdF8NWB3AVydv4xF_p13bw49JtN1PsF-jBdolm_lp-UKVulJreMezYOgkYH_f6Fh7R9BjiPgsPKFFHiNOhuLZTVWvIymM1xB55nUpVDM2VxluvbpKWyjCLRvBsFinally, I created a **logistic regression** model to predict whether an avocado was organic or conventional solely using average price. This model performed very well (~90% accuracy) considering it only takes one factor into account. This could be improved by adding more relevant variables.